**Data Frame, Files in R**

**Lab 3**

**R Data Frame**

* A data frame is a two-dimensional data structure which can store data in tabular format.
* Data frames have rows and columns and each column can be a different vector. And different vectors can be of different data types.
* DataFrame is made up of three principal components, the data, rows, and columns.

**Create Dataframe in R Programming Language**

# creating a data frame

friend.data<- data.frame(friend\_id = c(1:5),

friend\_name = c("Sachin", "Sourav","Dravid", "Sehwag","Dhoni"))

# print the data frame

print(friend.data)

**Output:**

friend\_idfriend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

**Summary of data in the data frame**

In R data frame, the statistical summary and nature of the data can be obtained by applyingsummary() function.

print(summary(friend.data))

**Output:**

friend\_idfriend\_name

Min.:1 Length:5

1st Qu.:2 Class:character

Median:3Mode:character

Mean:3

3rd Qu.:4

Max.:5

**Extract Data from Data Frame in R Language**

# Extractingfriend\_name column

result<- data.frame(friend.data$friend\_name)

print(result)

**Expand Data Frame**

# Expanding data frame

friend.data$location<- c("Kolkata", "Delhi","Bangalore", "Hyderabad","Chennai")

resultant<- friend.data

# print the modified data frame

print(resultant)

**Combine Data Frames**

In R, we use the rbind() and the cbind() function to combine two data frames together.

* rbind() - combines two data frames vertically
* cbind() - combines two data frames horizontally

**Example**

# create a data frame

dataframe1 <- data.frame (Name = c("Juan", "Alcaraz"),Age = c(22, 15))

# create another data frame

dataframe2 <- data.frame (Name = c("Yiruma", "Bach"),Age = c(46, 89))

# combine two data frames vertically

updated<- rbind(dataframe1, dataframe2)

print(updated)

Name Age

1 Juan 22

2 Alcaraz 15

3 Yiruma 46

4 Bach 89

**Use of which() function**

#interpret the results

> which(updated$Name=="Juan")

[1] 1

> updated[which(updated$Name=="Juan")]

Name

1 Juan

2 Alcaraz

3 Yiruma

4 Bach

> updated[which(updated$Name=="Juan"),]

Name Age

1 Juan 22

> updated[which(updated$Name=="Juan"),]$Age

[1] 22

>updated[which(updated$Name=="Juan"),]$Age=24

>

> updated

Name Age

1 Juan 24

2 Alcaraz 15

3 Yiruma 46

4 Bach 89

**Difference between data frame and lists**

# creating a data frame

> friend<- data.frame(friend\_id = c(1:5),friend\_name = c("Sachin", "Sourav","Dravid", "Sehwag","Dhoni"))

> print(friend)

friend\_id friend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

#creating a list from the same data

> friendlist <- list(friend\_id = c(1:5),friend\_name = c("Sachin", "Sourav","Dravid", "Sehwag","Dhoni"))

> # print the list

> print(friendlist)

$friend\_id

[1] 1 2 3 4 5

$friend\_name

[1] "Sachin" "Sourav" "Dravid" "Sehwag" "Dhoni"

#Types of data frame and list

> typeof(friend)

[1] "list"

> typeof(friendlist)

[1] "list"

> class(friend)

[1] "data.frame"

> class(friendlist)

[1] "list"

> dframe <- data.frame(friendlist)

> dframe

friend\_id friend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

> lframe <- list(dframe)

> lframe

[[1]]

friend\_id friend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

> dframe

friend\_id friend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

#size of data frame

> length(dframe)

[1] 2

> dframe <- data.frame(friendlist)

> dframe

friend\_id friend\_name

1 1 Sachin

2 2 Sourav

3 3 Dravid

4 4 Sehwag

5 5 Dhoni

> nrow(dframe)

[1] 5

**Accessing files for data mining**

file1 <- read.csv("college-perf.csv")

> file1

SAT GPA Projects Community Income Perf

1 1380 2.53 1 0 41800 Medium

2 1100 3.18 1 5 37600 High

3 1110 2.73 2 10 34800 Low

4 1180 2.49 3 0 24100 Medium

5 1240 2.89 3 5 56000 High

6 1140 2.85 2 0 50800 Low

…..

#class of file1

> typeof(file1)

[1] "list"

# Data is read as a data frame

> class(file1)

[1] "data.frame"

> length(file1)

[1] 6

> nrow(file1)

[1] 4000

> colnames(file1)

[1] "SAT" "GPA" "Projects" "Community" "Income" "Perf"

**Functions can be applied to the file data to extract meaning full information**

> mean(file1$SAT)

[1] 1192.84

> sum(file1$SAT)

[1] 4771360

> sum(which(file1$SAT > 100))

[1] 8002000

> max(file1$SAT)

[1] 1580